(12) UK Patent Application (19) GB (11) 2 250 405 (13) A

(43) Date of A publication 03.06.1992

(21) Application No 9119492.8

(22) Date of filing 11.09.1991

(30) Priority data
(31) 9019829 (32) 11.09.1990 (33) GB

(71) Applicant
British Telecommunications Public Limited Company

(Incorporated in the United Kingdom)

81 Newgate Street, London, EC1A 7AJ,
United Kingdom

(72) Inventor
Alison Diane Simons

(74) Agent and/or Address for Service
B G W Lloyd
Intellectual Property Unit, Room 1304,
151 Gower Street, London, WC1E 6BA,
United Kingdom

(51) INT CL[5]
G10L 3/00, G06F 15/72

(52) UK CL (Edition K)
H4T TCJA T126

(56) Documents cited
GB 2231246 A EP 0225729 A1 EP 0179701 A1
EP 0056507 A1

(58) Field of search
UK CL (Edition K) H4F FAA FDX FGG FGH FGS
FRX, H4T TBAX TCGD TCGX TCJA
INT CL[5] G10L 9/20
Online databases: WPI

(54) **Speech analysis and image synthesis**

(57) Successive frames of speech (at 1) are analysed 2, 3, 4 to produce a sequence of codewords identifying the character of the frame. A store 6 stores probability values P_{ei} indicating the probability that any codeword was produced by one of a set of standard 'mouth shapes', whilst a second store 7 stores values t_{ij} indicating the probability of one mouth shape following another. A Viterbi decoder examines the sequence of codewords and, using the probabilities, estimates the most likely sequence of mouth shapes to correspond to the speech. This can be used to generate a synthetic "talking face" moving image, e.g. for videophone or audio conferencing applications.

Fig.1.

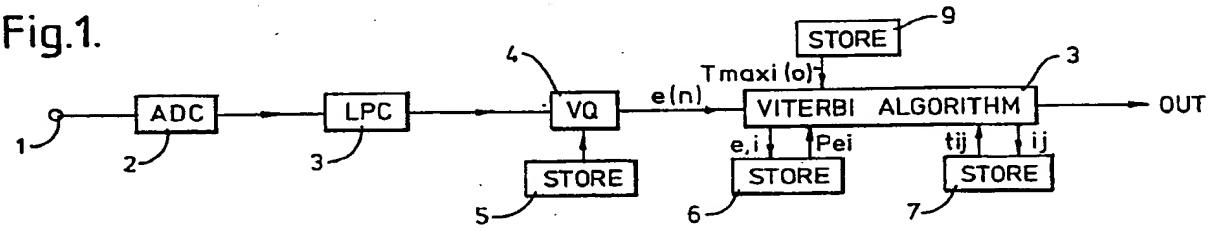
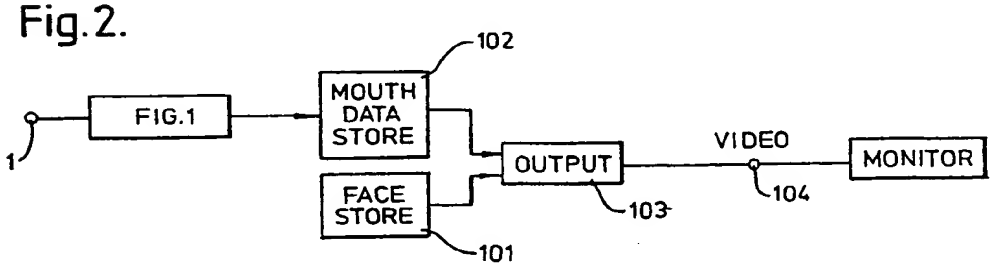


Fig.2.



At least one drawing originally filed was informal and the print reproduced here is taken from a later filed formal copy.

GB 2 250 405 A

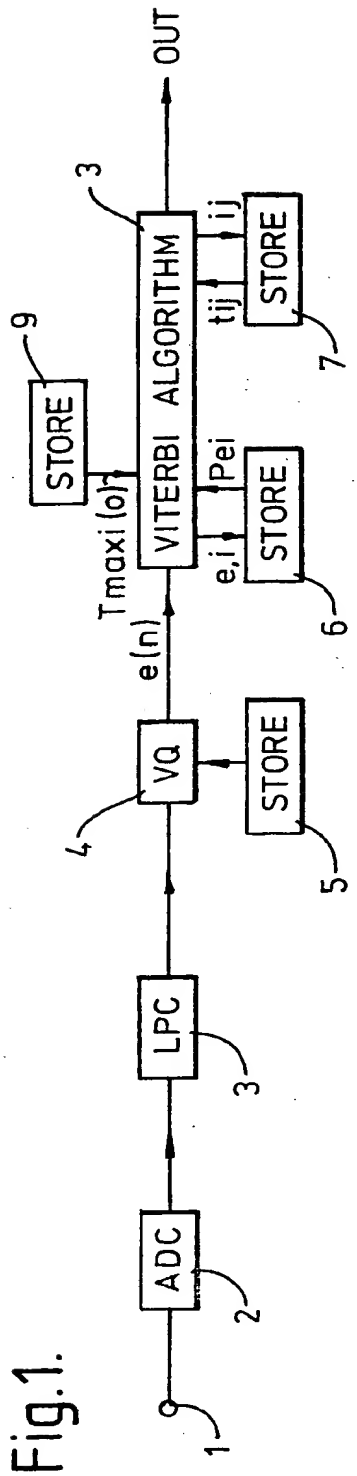

OUT
3
VITERBI ALGORITHM
tij
ij
STORE
7
9
STORE
Pei
Tmaxi (o)
e,i
STORE
6
e(n)
VQ
STORE
4
5
LPC
3
ADC
2
1

Fig.1.

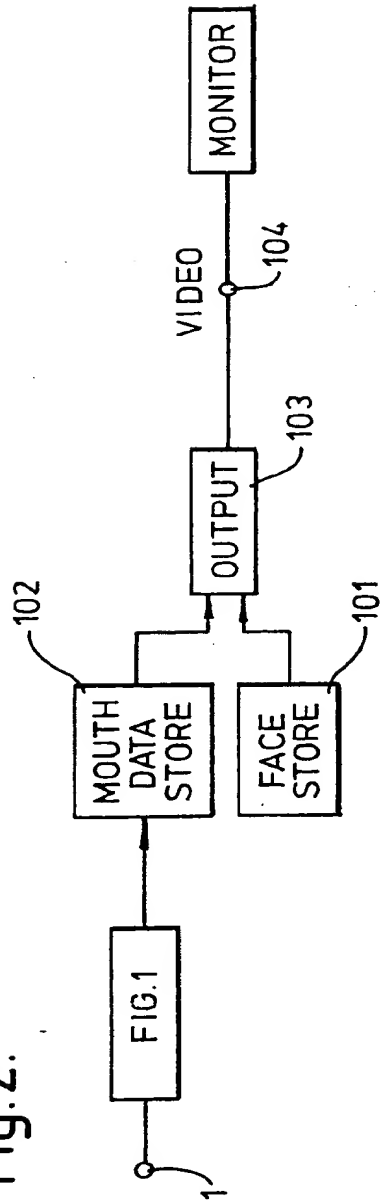

MONITOR
VIDEO
104
OUTPUT
103
102
101
MOUTH DATA STORE
FACE STORE
FIG.1
1

Fig.2.

2230405

SPEECH ANALYSIS AND IMAGE SYNTHESIS

The present application relates to the analysis of speech, and more particularly to the analysis of speech to estimate the visual appearance of a mouth by which the speech is uttered. One specific application of such analysis is for the synthesis on the basis of an input speech signal, of a moving image of a human face for display to accompany such speech. Such synthesis may be desired for a video terminal in low bit rate transmission systems, or for enhanced audio-conferencing facilities.

In our European patent application no. 86308732.6 (Publication No. 0225729A) we describe an apparatus for synthesis of a moving image which has a store for an image of a face, and a store for storing a set of data blocks each corresponding to the mouth area of the face and representing a respective different mouth shape. In operation, an input audio signal is analysed to produce sequences of spectral parameters which are then used to access a table relating these parameters to codewords identifying mouth data blocks, the codewords obtained being employed to select the corresponding mouth data blocks for output to a display device.

The present invention is defined in the claims.

Some embodiments of the invention will now be described with reference to the accompanying drawings in which:

Figure 1 is a block diagram of one form of speech analysis apparatus in accordance with the invention; and

Figure 2 is a schematic diagram illustrating the operation of the apparatus; and

Figure 3 is an apparatus for synthesising a moving image, incorporating the speech analysis apparatus of Figure 1.

The purpose of the speech analysis apparatus shown in Figure 1 is to receive an input speech signal at an input 1, analyse it, and estimate at intervals the mouth shape which was most likely to have produced that portion of speech. The output of the apparatus is a sequence of codewords each of which identifies an entry in a codebook of mouth shapes. In this example, the codebook is assumed to contain 16 entries. The actual mouth shapes are not stored in the apparatus of Figure 1.

In the embodiment shown, the speech is firstly sampled at 8kHz and converted into digital form in an analogue-to-digital converter 2 and processed by an LPC analysis unit 3 to produce for successive frames of speech (e.g. of 20ms duration) a set of eight LPC coefficients defining a filter having a spectral response similar to that of the speech frame. Any of the conventional LPC analysis methods commonly used for LPC speech coders may be employed for this purpose. The coefficients are vector quantised in a VQ unit 4, which matches each set of coefficients to the nearest entry in a codebook of (e.g. sixty-four) coefficient sets stored in a speech codebook store 5. This process is again conventional; for example the entry chosen may be that for which the City Block distance (viz the sum of the moduli of the intercoefficient differences) between the actual set and the stored set is a minimum.

The use of vector quantised LPC coefficients is one possible example; LPC cepstral coefficients, or extraction of other speech features, as is common in speech recognition systems, may alternatively be employed.

The apparatus also includes a store 6 containing probability values p_{ei} each of which indicates the relative probability that the speech represented by codeword e originated from a mouth having the shape represented by codeword i. In this example, the store has 16 x 64 = 1024 entries.

A further store 7 contains transition probability values t_{ij} each of which indicates the relative probability that mouth shape i is followed by mouth shape j; thus it has 16 x 16 = 256 entries.

The mouth codebook, the speech codebook, and the probability values are generated by a training process, analysing a video signal and its accompanying speech. In a test, fifty sentences totalling 200 seconds were used thereby generating 10,000 frames of speech data and (at 25 frames/second) 5,000 frames of video data.

The speech codebook is generated by selecting that set of 64 coefficient sets which substantially minimises the average distance between a training frame and the nearest entry in the set. Similarly, the mouth codebook was generated by selecting that set of sixteen mouth shapes which substantially minimises the average distance between mouth portions of a training video frame and the nearest entry in the set. The distance measure used was simply the City Block distance between two mouths as represented by the height and width of the mouth opening, but naturally more sophisticated measures could be used if desired.

The mouth/speech probability values p_{ei} are generated by, for each frame, matching the video and speech frames to the nearest codebook entry; the number of occurences of each pair is recorded. For this part of the training, each video frame was repeated once. Likewise the transitional mouth probability values t_{ij} are

obtained by counting the number of occurences of mouths i and j appearing consecutively.

If any particular event did not occur in the training sequence, the corresponding probability value was set to a small value rather than zero.

Returning to Figure 1, the apparatus includes a Viterbi algorithm unit 8. The purpose of this is, for a passage of speech containing N frames, to determine the most likely sequence of mouth shapes having regard to (1) the observed speech as represented by the codewords e(n) (n=1 ... N) output by the VQ unit 4 (2) the speech/mouth shape probability values p_{ei} stored in the store 6 and (3) the transitional probability values t_{ij} stored in the store 7. This process is, illustrated schematically in Figure 2.

The application of the Viterbi algorithm to the analysis of the speech information will now be described. For an utterance (or other portion) of speech a sequence of speech codewords, have been produced, one for each frame of speech. The basic procedure is that one calculates, for each frame in succession, the probability that that frame resulted from each of the permitted mouth shapes, taking into account the speech codeword for that frame, the calculated probability for the preceding frame, and the stored probability values. When the end of the sentence is reached, the mouth shape associated with the largest of the calculated probabilities is chosen for that speech frame, whereupon one then re-visits successive preceding frames and makes a similar decision taking into account the previous decision (in respect of the following frame).

We recall that the probability that a particular speech frame having codeword e was generated by mouth shape i is p_{ei}, and that the probability of mouth shape

i being followed by shape j is t_{ij} - these values being stored in the stores 6, 7. We define $P_i(n)$ for the nth frame as being the calculated probability that that frame resulted from mouth i.

We commence by finding $P_i(1)$ (i = 0 ... 15) for the first frame. There is no previous frame, so we estimate these probabilities on the basis of the codeword e(1) for that frame. Thus:

$P_i(1) = p_{e(1)i}$ (i = 0 ... 15).

For the second frame, we first apply the stored transitional probability values t_{ij} to the calculated probabilities $P_i(1)$ from frame 1. For each candidate second frame mouth shape, we multiply $P_i(1)$ by the corresponding transitional probabilities - for the jth shape:

$T_{ij}(2) = P_i(1)t_{ij}$ (i = 0 ... 15)

Now select the largest of these, $T_{max\ j}(2)$ noting also the value of i associated with it.

(= $i_{max}(2,j)$)

Note that the significance of this is that <u>if</u> mouthshape j is the shape chosen for frame 2 then shape $i_{max}(2,j)$ is the most likely one to precede it in frame 1 having regard to the calculated probabilities $P_i(1)$ and the transitional probability values t_{ij}.

Having found all these maxima $T_{max\ i}(2)$ (i = 0 ... 15) we then use the frame 2 codeword e(2) to obtain the probability values $p_{e(2)i}$ from the store and multiply the two to obtain.

$P_i(2) = T_{max\ i}(2) \cdot p_{e(2)i}$

- the calculated probabilities for the second frame.

This process is repeated for successive frames, until we have found $P_i(N)$ for the last frame of the utterance. At this point the first actual decision is made; mouth shape I(N) associated with the largest of the

set $P_i(N)$ being chosen, where I(N) is the associated value of i.

Recalling that each time we selected the maximum value of $T_{ij}(n)$ we recorded the previous frame mouth shape $i_{max}(n-1,j)$, associated with the choice, we can now go back to the penultimate frame N-1 and deduce that this implies selection of shape $i_{max}(N-1, I(N))$ for it.

In the above description, the calculated probability value for the first frame was estimated simply on the basis of the received speech, since there was no previous frame. In a modification, a further store 9 is included which contains the probability for each mouth shape of its occuring at the beginning of an utterance. These values are then used for the first frame just as the product $T_{max\ i}$ is used for later frames.

This description assumes that the application of the Viterbi algorithm is performed after the whole of an utterance (e.g. a sentence) has been received. Where it is desired to perform the analysis in real time this may involve an undesirable delay, and a modified approach may be preferred. In that the algorithm relies upon the history of the speech over a period, some delay is inherent; however tests indicate that traceback over a period greater than 200ms is of little value.

In the modified method, assuming operation over a window of m frames in length, suppose we start analysis at frame n, the mouth shape for frame n-1 having already been fixed (unless n = 1 in which case that frame is dealt with as already discussed for the first frame). As frame n-1 is fixed, the previous frame probability $P_i(n-1)$ is unity for the selected entry and zero for all other values of i. The above procedure is then followed up to frame n+m, the mouth codeword for which has just been generated by the VQ unit 4, with traceback to frame n. The

mouthshape for frame n is then fixed, and the decisions for later frames discarded. When frame n+m+1 is available, the process is the repeated starting at frame n+1 and extending up to frame n+m+1.

Figure 2 shows an apparatus for synthesis of a moving image including a human face having a mouth which moves in correspondence with a speech signal received at an input 1. The input 1 is connected to the input of an analysis apparatus 100 which is of the structure described above and shown in Figure 1.

A store 101 contains data representing a stored image of a face; this may be for example a stored digital representation of a raster-scan television picture, the store being a conventional video frame store.

A second store 102 stores sixteen data blocks each being a digital representation of a mouth having a respective one of the mouth shapes discussed above.

For the purposes of the present description it is assumed that each block is stored in pixel map form - i.e. a mouth could be superimposed into the face stored in the store 101 simply by writing each picture element from the block into the appropriate location in the store 101. However, other, parametric, methods of mouth representation could be used.

The analysis apparatus 100 produces at its output, every 20ms, a codeword identifying one of the data blocks in the store 102. An output unit 103 reads the face data from the store 101 every 20ms to form a raster scan television signal; when however it requires picture data in respect of a portion of the picture area corresponding to the mouth, it reads the data instead from the portion of the mouth data store 102 identified by the codeword supplied by the analysis apparatus 100, so that the desired mouth shape is incorporated into the image. The

video signal output on an output line 104 can be displayed on a video monitor 105.

This description assumes a video rate of 50 frames per second; in order to generate a signal at the UK Standard (System I) rate of 25 frames per second one could use a 40ms speech frame period, but if it is preferred to retain the 20ms analysis period then the frame rate could be reduced by omitting alternate mouth shapes, or (more preferably to avoid aliasing) temporally filtering the mouth image sequence and then discarding alternate images. Obviously other video frame rates such as 30 frames/second used in system M can be achieved by similar such adjustments.

CLAIMS

1. A speech analysis apparatus comprising means for analysing a speech signal to generate at intervals items of speech information each representative of the speech during that interval;

first store means storing, for each possible item of speech information, data representing the probabilities that a particular portion of speech corresponding to that item of information has been generated by a mouth having each of a predetermined plurality of mouth shapes;

second store means storing, for each possible one of the plurality of mouth shapes, data representing the probabilities that that shape is followed by that shape and by each other one of those shapes; and

decoding means responsive to the generated items of speech information, to the probability data stored in said first store means in respect of those generated items, and to the probability data stored in the second store means to determine a sequence of mouth shapes deemed to be substantially most likely to correspond to the said generated items.

2. An apparatus according to claim 1 in which the speech analysis means comprises

(a) means to analyse each of successive frames of speech to produce therefor a set of parameters representing the spectral content thereof;

(b) a fourth store storing a plurality of reference sets of parameters; and

(c) means to determine, for each produced parameter set, which of the stored sets it most closely resembles;

the said items of speech information being codewords identifying the determined reference sets.

3. An apparatus according to claim 2 in which the parameters are the coefficients of a linear prediction filter.

4. An apparatus according to claim 1, 2 or 3 in which the decoding means is a Viterbi decoder.

5. An apparatus according to any one of the preceding claims including further store means storing for each of the plurality of mouth shapes data representing the probability of its occurence at the commencement of an utterance, the decoding means being responsive also to the data stored in the further store means.

6. An apparatus for synthesis of a moving image, comprising

(a) means for storage and output of data representing an image of a face;

(b) means for storage and output of a set of mouth data blocks each representing an image of a mouth having a respective one of the said shapes;

(c) a speech signal input;

(d) a speech analysis apparatus according to any one of the preceding claims to receive the speech input; and

(e) control means responsive to the output of the speech analysis apparatus to select for output from the mouth data storage means the data blocks corresponding to the determined sequence.

7. An apparatus according to claim 6 further including video signal generating means operable to generate video frames each representing a face image corresponding to the stored face data having superimposed thereon a mouth image represented by a said selected mouth data block.

8. An apparatus for speech analysis substantially as herein described with reference to the accompanying drawings.

9. An apparatus for synthesis of a moving image substantially as herein described with reference to the accompanying drawings.

Patents Act 1977
Examiner's report to the Comptroller under Section 17 (The Search Report)

Application number

9119492.8

Relevant Technical fields

(i) UK Cl (Edition K) H4F (FAA, FDX, FRX, FGG, FGH, FGS) H4T (TBAX, TCGD, TCGX, TCJA)

(ii) Int Cl (Edition 5) G10L 9/20

Search Examiner

P J EASTERFIELD

Databases (see over)

(i) UK Patent Office

(ii) ONLINE DATABASES: WPI

Date of Search

18 FEBRUARY 1992

Documents considered relevant following a search in respect of claims 1-7

Category (see over)	Identity of document and relevant passages	Relevant to claim(s)
A	GB 2231246 A (DENWA)	
A	EP 0225729 A1 (BT)	
A	EP 0179701 A1 (GUINET)	
A	EP 0056507 A1 (BLOOMSTEIN)	

SF2(p)

HD - c:\wp51\doc99\fil000067

Category	Identity of document and relevant passages	Relevant to claim(s)

Categories of documents

X: Document indicating lack of novelty or of inventive step.

Y: Document indicating lack of inventive step if combined with one or more other documents of the same category.

A: Document indicating technological background and/or state of the art.

P: Document published on or after the declared priority date but before the filing date of the present application.

E: Patent document published on or after, but with priority date earlier than, the filing date of the present application.

&: Member of the same patent family, corresponding document.

Databases: The UK Patent Office database comprises classified collections of GB, EP, WO and US patent specifications as outlined periodically in the Official Journal (Patents). The on-line databases considered for search are also listed periodically in the Official Journal (Patents).

Published 1992 at The Patent Office, Concept House, Cardiff Road, Newport, Gwent NP9 1RH. Further copies may be obtained from Sales Branch, Unit 6, Nine Mile Point, Cwmfelinfach, Cross Keys, Newport, NP1 7HZ. Printed by Multiplex techniques ltd, St Mary Cray, Kent.